# Practical LLM projects at Te Herenga Waka

# Five projects exploring practical use of LLMs at Vic

New AI research tools available in the Library

- Marcus Harvey (Library)

Vic's Working Group on AI for teaching and learning

- Stella McIntosh (Acad. Integrity), Rob Stratford (Acad. Office)

Pūaha AI: a student-facing AI chatbot offering advice

- Leanne Gibson (CIO), Yvonne Green (Project Manager)

A staff-facing policy chatbot project

- Matt Farrington (VUW Legal Counsel)

Advice to government departments on LLMs, from the Policy Hub

- Andrew Jackson (Policy Hub), Simon McCallum (ECS)

# Pūaha AI

# Pūaha AI

Who's doing this?

- Coordinated by Vic's Digital Services Team
- Contractor: Fusion5
- Microsoft are donating staff time.

# Pūaha AI

Who's doing this?

- Coordinated by Vic's Digital Services Team
- Contractor: Fusion5
- Microsoft are donating staff time.

What's being built?

- A chatbot that runs on Pūaha, Vic's student self-service portal.

# Pūaha AI

Who's doing this?

- Coordinated by Vic's Digital Services Team
- Contractor: Fusion5
- Microsoft are donating staff time.

What's being built?

- A chatbot that runs on Pūaha, Vic's student self-service portal.
- Pūaha is 'where you can find tools & information for your studies [including]
  - timetable, grades, student records, and other academic details
  - key dates, links to news and events
  - links to services that support your learning, study, wellbeing'

# System design

# System design

They're using Microsoft's Azure AI Studio:

- This is a programmer's interface to OpenAI's LLMs.
- These LLMs are OpenAI ChatGPT models.

# System design

They're using Microsoft's Azure AI Studio:

- This is a programmer's interface to OpenAI's LLMs.
- These LLMs are OpenAI ChatGPT models.

In AI studio, they're using Copilot Studio:

- A 'low-code development tool' that extends MS copilot's functionality.
- Lets you create new gen AI assistants, or enhance existing ones.

# System design

They're using Microsoft's Azure AI Studio:

- This is a programmer's interface to OpenAI's LLMs.
- These LLMs are OpenAI ChatGPT models.

In AI studio, they're using Copilot Studio:

- A 'low-code development tool' that extends MS copilot's functionality.
- Lets you create new gen AI assistants, or enhance existing ones.

The basic idea is to fine-tune an OpenAI LLM with some Vic-specific text, and limit the LLM to speak about this same text.

- The text is the Pūaha text base.

# System design

They're using Microsoft's Azure AI Studio:

- This is a programmer's interface to OpenAI's LLMs.
- These LLMs are OpenAI ChatGPT models.

In AI studio, they're using Copilot Studio:

- A 'low-code development tool' that extends MS copilot's functionality.
- Lets you create new gen AI assistants, or enhance existing ones.

The basic idea is to fine-tune an OpenAI LLM with some Vic-specific text, and limit the LLM to speak about this same text.

- The text is the Pūaha text base.
- There's also some manual conversation scripting in there.

# Design process

# Design process

They ran a *pilot* system in Week 0 of this Trimester.

- It ran in a booth, with constant supervision.

# Design process

They ran a *pilot* system in Week 0 of this Trimester.

- It ran in a booth, with constant supervision.
- There was a preliminary evaluation.
- This went well enough to contract a full project, to go live in T2.

# Design process

They ran a *pilot* system in Week 0 of this Trimester.

- It ran in a booth, with constant supervision.
- There was a preliminary evaluation.
- This went well enough to contract a full project, to go live in T2.

The process for the full project:

# Design process

They ran a *pilot* system in Week 0 of this Trimester.

- It ran in a booth, with constant supervision.
- There was a preliminary evaluation.
- This went well enough to contract a full project, to go live in T2.

The process for the full project:

- Several workshops, with a range of participants

# Design process

They ran a *pilot* system in Week 0 of this Trimester.

- It ran in a booth, with constant supervision.
- There was a preliminary evaluation.
- This went well enough to contract a full project, to go live in T2.

The process for the full project:

- Several workshops, with a range of participants
- An ethics/risk review meeting (which I asked for)

# Design process

They ran a *pilot* system in Week 0 of this Trimester.

- It ran in a booth, with constant supervision.
- There was a preliminary evaluation.
- This went well enough to contract a full project, to go live in T2.

The process for the full project:

- Several workshops, with a range of participants
- An ethics/risk review meeting (which I asked for)
- A meeting about evaluation procedures

# Design process

They ran a *pilot* system in Week 0 of this Trimester.

- It ran in a booth, with constant supervision.
- There was a preliminary evaluation.
- This went well enough to contract a full project, to go live in T2.

The process for the full project:

- Several workshops, with a range of participants
- An ethics/risk review meeting (which I asked for)
- A meeting about evaluation procedures
- A meeting with some MS people, who can tell us more about how system reliability is assured.

# Design process

They ran a *pilot* system in Week 0 of this Trimester.

- It ran in a booth, with constant supervision.
- There was a preliminary evaluation.
- This went well enough to contract a full project, to go live in T2.

The process for the full project:

- Several workshops, with a range of participants
- An ethics/risk review meeting (which I asked for)
- A meeting about evaluation procedures
- A meeting with some MS people, who can tell us more about how system reliability is assured.
  - Side note: OpenAI's Head of Alignment, Jan Leike, just quit: 'Over the past years, safety culture and processes have taken a backseat to shiny products. . . '

# Evaluation processes

# Evaluation processes

Testing will be done by:

- The Applications Support team (at Digital Solutions)
- Vic's Student Service Centre (with a staff of student advisors)
- Simon and me

# Evaluation processes

Testing will be done by:

- The Applications Support team (at Digital Solutions)
- Vic's Student Service Centre (with a staff of student advisors)
- Simon and me

What will be done?

# Evaluation processes

Testing will be done by:

- The Applications Support team (at Digital Solutions)
- Vic's Student Service Centre (with a staff of student advisors)
- Simon and me

What will be done?

- Several different 'user profiles' (including 'hostile' users)

# Evaluation processes

Testing will be done by:

- The Applications Support team (at Digital Solutions)
- Vic's Student Service Centre (with a staff of student advisors)
- Simon and me

What will be done?

- Several different 'user profiles' (including 'hostile' users)
- Several quantitative measures of system performance
  - Including accuracy, friendliness, 'reputational risk'...

# Evaluation processes

Testing will be done by:

- The Applications Support team (at Digital Solutions)
- Vic's Student Service Centre (with a staff of student advisors)
- Simon and me

What will be done?

- Several different 'user profiles' (including 'hostile' users)
- Several quantitative measures of system performance
  - Including accuracy, friendliness, 'reputational risk'...
- Interactions logged, and annotated (by multiple annotators?)

# Evaluation processes

Testing will be done by:

- The Applications Support team (at Digital Solutions)
- Vic's Student Service Centre (with a staff of student advisors)
- Simon and me

What will be done?

- Several different 'user profiles' (including 'hostile' users)
- Several quantitative measures of system performance
  - Including accuracy, friendliness, 'reputational risk'...
- Interactions logged, and annotated (by multiple annotators?)
- Comparisons against logged human-human interactions?

# Evaluation processes

Testing will be done by:

- The Applications Support team (at Digital Solutions)
- Vic's Student Service Centre (with a staff of student advisors)
- Simon and me

What will be done?

- Several different 'user profiles' (including 'hostile' users)
- Several quantitative measures of system performance
  - Including accuracy, friendliness, 'reputational risk'...
- Interactions logged, and annotated (by multiple annotators?)
- Comparisons against logged human-human interactions?
- Checking of MS's own red-teaming methods